Analysis of Integrated Data without Data Integration

Alan F. Karr, Xiaodong Lin, Ashish P. Sanil
National Institute of Statistical Sciences
Research Triangle Park, NC 27709–4006, USA
karr@niss.org, linxd@samsi.info, ashish@niss.org

Jerome P. Reiter
Duke University
Durham, NC 27708 USA
jerry@stat.duke.edu

April 5, 2004

1 Introduction

Many scientific and policy investigations require statistical analyses that "integrate" data stored in multiple, distributed databases. For example, a regression analysis on integrated state databases about factors influencing student performance would be more insightful than individual analyses, or at least complementary to them. Other contexts where the same need arises range from homeland security to environmental monitoring.

At the same time, the barriers to actually integrating the databases are numerous. One is confidentiality: the database holders—we term them "agencies"—almost always wish to protect the identities of their data subjects. Another is regulation: the agencies may be forbidden by law to share their data, either with each other or with a trusted third party. A third is scale: despite advances in networking technology, the only way to move a terabyte from point A today to point B tomorrow is FedEx.

The good news is that for many analyses it is not even necessary to move the data. Instead, using techniques from computer science known generically as secure multi-party computation, the agencies can share summaries of the data anonymously, but in a way that the analysis can be performed in a statistically sound manner.

We illustrate in this paper for linear regression on "horizontally partitioned data." Only one concept is needed, that of secure summation, which is described in §3. There are both other approaches to this problem and similar approaches to related problems, such as vertically partitioned data. There are also alternative approaches for lower risk situations. For example, we have devel-

oped techniques for secure data integration, which build the integrated database in such a way that no agency can determine the source of any data elements other than its own.

2 The Problem

We assume that there are $K > 2$ agencies, each with the same numerical data on its own n_j data subjects—p predictors X^j and a response y^j, and that the agencies wish to fit the usual linear model

$$y = X\beta + \epsilon, \tag{1}$$

to the "global" data

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} y^1 \\ \vdots \\ y^K \end{bmatrix}. \tag{2}$$

Figure 2 shows this horizontal partitioning for $K = 3$ agencies. Each X^j is $n_j \times p$.

Under the condition that $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, the least squares estimator for β is of course

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{3}$$

We show in §4 how $\hat{\beta}$ can be computed without integrating the agencies' databases.

Several assumptions about agency behavior are necessary. First, the agencies agree to cooperate to perform the regression, and none of them is specifically interested in breaking the confidentiality of any of the others' data. Second, each reports accurately the results of computations on its own data, and follows the agreed-on computational protocols, such as secure summation, properly. And finally, there is no collusion among agencies.

3 Secure Summation

The simplest secure multi-party computation, and essentially the only one needed for secure regression, is to sum values v_j held by the agencies. Let v denote the sum. The method described below,[1] lets agency j learn only the minimum possible about the other agencies' values, namely, the sum $v_{(-j)} = \sum_{\ell \neq j} v_\ell$.

The secure summation protocol, which is shown pictorially in Figure 1, is almost more complicated to describe than to implement. Number the agencies $1, \ldots, K$. Agency 1 generates a very large random integer R, adds R to its value v_1, and sends the sum to agency 2. Since R is random, Agency 2 learns effectively nothing about v_1. Agency 2 adds its value v_2 to $R + v_1$, sends the result to agency 3, and so on. Finally, agency 1 receives $R + v_1 + \ldots + v_K = R + v$ from agency K, subtracts R, and shares the result v with the other agencies. Here is one place where cooperation matters: agency 1 is obliged to share v with the other agencies.

[1] Which has appeared recently in the puzzles of the radio shows *Car Talk* and *NPR Weekend Edition Sunday*.

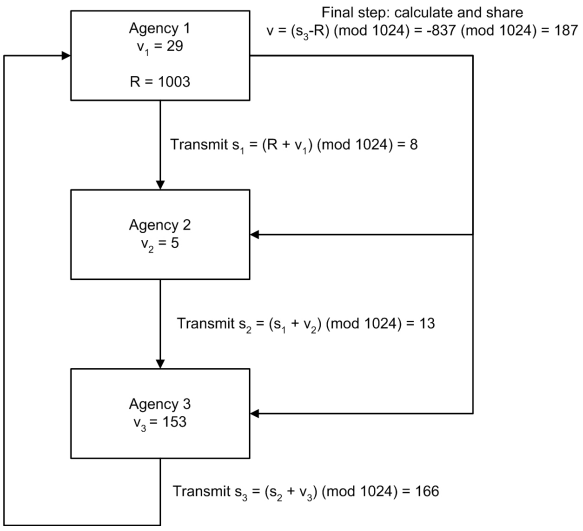

Figure 1: Values computed at each agency during secure computation of a sum initiated by Agency 1. Here $v_1 = 29$, $v_2 = 5$, $v_3 = 152$ and $v = 187$. All arithmetic is modulo $m = 1024$.

An extra layer of protection is possible, as shown in Figure 1. Suppose that v is known to lie in the range $[0, m)$, where m is a very large number, say 2^{100}, known to all the agencies. Then R can be chosen randomly from $\{0, \ldots, m-1\}$ and all computations performed modulo m.

Here is a simple application: the agencies have income data and wish to compute the global average income. Let n_j be the number of records in agency j's database and I_j be the sum of their incomes. The quantity to be computed is $\bar{I} = \sum_j I_j / \sum_j n_j$. But this is easy: the numerator can be computed using secure summation on the I_j's, and the denominator can be computed using secure summation on the n_j's.

4 Secure Regression

To compute $\hat{\beta}$ in (3), it is necessary to compute $X^T X$ and $X^T y$. Because of the partitioning in (2),

$$X^T X = \sum_{j=1}^{K} (X^j)^T X^j.$$

So agency j computes its own $(X^j)^T X^j$, which has dimensions $p \times p$, where p is the number of data attributes, and these are combined entry-wise using secure summation. This computation is illustrated pictorially with $K = 3$ in Figure 2. Similarly, since by (2)

$$X^T y = \sum_{j=1}^{K} (X^j)^T y^j,$$

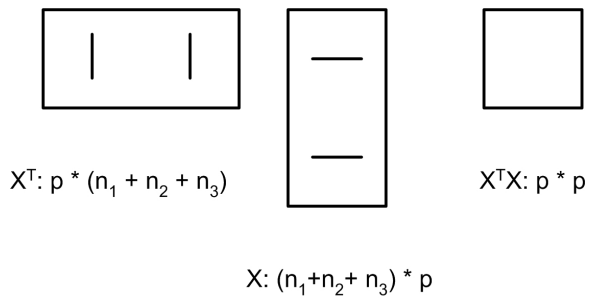

Figure 2: Pictorial representation of the secure regression computation in §4. The dimensions of various matrices are shown.

$X^T y$ can be computed by secure, entry-wise summation of the $(X^j)^T y^j$.

Finally, each agency can calculate $\hat{\beta}$ from the shared values of $X^T X$ and $X^T y$ using (3). Note that no agency learns any other agency's $(X^j)^T X^j$ or $(X^j)^T y^j$, but only the sum of these over all the other agencies.

5 Model Diagnostics

In the absence of model diagnostics, secure regression as described in §4 loses much of its appeal, especially to statisticians. We describe briefly two strategies for producing informative diagnostics. The first is to use diagnostics that can be computed using secure summation from corresponding local statistics. The second uses "secure data integration" (Karr et al., 2004) to share synthetic residuals (Reiter, 2003).

Among diagnostics computable by secure summation are the coefficient of determination R^2, the least squares estimate $S^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})/(n - p)$ of the error variance σ^2, correlations between predictors and residuals, and the hat matrix $H = X(X^T X)^{-1} X^T$, which can be used to identify X-outliers.

For diagnosing some types of assumption violations, only patterns in relationships among the residuals and predictors suggestive of model mis-specification are needed, rather than exact values of the residuals and predictors. Such diagnostics can be produced for the global database using secure data integration protocols (Karr et al., 2004) to share synthetic diagnostics proposed for remote access computer servers (Gomatam et al., 2003).

The synthetic diagnostics are generated in three steps. First, each agency simulates values of its predictors. Second, using the global regression coefficients, each agency simulates residuals associated with these synthetic predictors in a way—and this is the hard part—that mimics the relationships between the predictors and residuals in its own data. Finally, the agencies share their synthetic predictors and residuals using secure data integration.

6 Discussion

In this paper we have presented a framework for secure linear regression in a cooperative environment. A huge number of variations is possible. For example, in order to give the agencies flexibility, it may be important to give them the option of withdrawing from the computation when their perceived risk becomes too great. To illustrate, agency j may wish to withdraw if its sample size n_j is too large relative to the global sample size n. This is the classical p-rule in the statistical disclosure limitation literature (Willenborg and de Waal, 2001). But, as noted in §3, n can be computed using secure summation, and so agencies may then "opt out" according to whatever criteria they wish to employ. It is even possible, at least under a scenario that the process does not proceed if any of the agencies opts out, to allow the opting out itself to be anonymous.

Acknowledgements

This research was supported by NSF grant EIA–0131884 to the National Institute of Statistical Sciences.

References

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2003). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.* Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004). Secure regression on distributed databases. *J. Computational and Graphical Statist.* Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.

Reiter, J. P. (2003). Model diagnostics for remote access regression servers. *Statistics and Computing*, 13:371–380.

Willenborg, L. C. R. J. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer–Verlag, New York.